02-28-00

A

02/25/00 JC670 U.S. PTO

PATENT APPLICATION
Attorney's Do. No. 6647-3

JC511 U.S. PTO 09/512963 02/25/00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

EXPRESS MAIL MAILING LABEL NO. EL253054261US
DATE OF DEPOSIT: FEBRUARY 25, 2000

I HEREBY CERTIFY THAT THIS PAPER AND ENCLOSURES AND/OR FEE ARE BEING DEPOSITED WITH THE UNITED STATES POSTAL SERVICE "EXPRESS MAIL POST OFFICE TO ADDRESSEE" SERVICE UNDER 37 CFR 1.10 ON THE DATE INDICATED ABOVE AND IS ADDRESSED TO: BOX PATENT APPLICATION, ASSISTANT COMMISSIONER FOR PATENTS, WASHINGTON D.C. 20231.

AMANDA HALE-WISENER
(SENDER'S PRINTED NAME)

Amanda Hale-Wisener
(SIGNATURE)

Box Patent Application
Assistant Commissioner for Patents
Washington, D.C. 20231

Enclosed for filing is a patent application under 37 CFR 1.53(b) of:

Inventors[or Application Identifier]: Delos C. Jensen and Stephen R. Carter

For: CONSTRUCTION, MANIPULATION, AND COMPARISON OF A MULTI-DIMENSIONAL SEMANTIC SPACE

[If continuing application] This application is a ☐ continuation, ☐ divisional, ☐ continuation-in-part of prior application Serial No. _____, filed _____.

Enclosures:
- ☒ Specification (pages 1-23); claims (pages 24-28); abstract (page 29)
- ☒ 21 sheet(s) of drawings
- ☒ Declaration or Combined Declaration and Power of Attorney
  - ☒ Newly executed (original or copy)
  - ☐ Copy from a prior application (37 CFR 1.63(d))
  - ☐ Incorporation by Reference--The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.
  - ☐ Deletion of Inventors (signed statement attached deleting inventor(s) named in the prior application (37 CFR 1.63(d)(2) and 1.33(b)
- ☒ Power of Attorney

☒ Assignment with cover sheet
☐ Certified copy of priority document:
☐ Information Disclosure Statement with Form PTO 1449
☐ Copies of references listed on attached Form PTO-1449
☐ Preliminary Amendment
☐ Change of Address

CLAIMS AS FILED				
For	Number Filed	Number Extra	Rate	Basic Fee $760.00
Total Claims	22-20	2	x $ 18 =	36.00
Independent Claims	8-3	5	x $ 78 =	390.00
Multiple Dependent Claim Fee			x $260 =	
TOTAL FILING FEE				$1,186.00

☐ Cancel in this divisional application original claims _____ of the prior application Serial No. _____ before calculating the filing fee. (At least one original independent claim must be retained for filing purposes.)

☒ A check in the amount of $1,226.00 to cover ☒ filing fee and ☒ assignment recordal fee ($40) is enclosed.

☒ Any deficiency or overpayment should be charged or credited to deposit account number 13-1703. A duplicate copy of this sheet is enclosed.

Customer No. 20575

Respectfully submitted,

MARGER JOHNSON & McCOLLOM, P.C.

Alexander C. Johnson, Jr.
Reg. No. 29,396

MARGER JOHNSON & McCOLLOM, P.C.
1030 S.W. Morrison Street
Portland, Oregon 97205
(503) 222-3613

CONSTRUCTION, MANIPULATION, AND COMPARISON OF A MULTI-DIMENSIONAL SEMANTIC SPACE

This application is related to co-pending U.S. Patent application Serial No. 09/109,804, titled "METHOD AND APPARATUS FOR SEMANTIC CHARACTERIZATION," filed July 2, 1998.

FIELD OF THE INVENTION

This invention pertains to determining the semantic content of a network, and more particularly to improving searching of the network.

BACKGROUND OF THE INVENTION

The Internet is about *content*. Content being accessed, published, indexed, analyzed, secured, purchased, stolen, vandalized, etc. Whether the content is white-papers, on-line books, catalogs, real-time games, address books, streaming audio and video, etc., it is content that people and cyber-agents are seeking. The future of the Internet lies not in bandwidth or capacity, but rather the ability to retrieve relevant content. Technology that allows fast and accurate access to relevant content will be used by the masses of carbon and silicon Internet users. Not because it is a better mouse-trap, but because controlled access to relevant content will allow the Internet to thrive, survive, and continue it's explosive growth. Fast and accurate semantic access to Internet content will determine who rules the next Internet era.

Caught between the sheer (and ever growing) volume of content, the huge and rapidly increasing number of Internet users, and a growing sophistication in the demands of those users, the current TCP/IP infrastructure and architecture is showing its inadequacies – it is a victim of its own success. One of the many strategies under consideration by the Internet community for redressing these inadequacies is to *build intelligence* into the network. Directory Services and Caching are two prime examples of intelligent network components. Adaptive routing with route caching is another example of an intelligent network component.

Yet another example of network intelligence that is receiving close attention these days is the characterization of content by its meaning (semantics). The obvious advantages that accrue with even a moderately successful semantic characterization component are such that almost everyone is tempted to dip a toe in the water. But assigning semantics to information on the Internet is the kind of undertaking that consumes vast amounts of resources.

Accordingly, a need remains for a way to assign semantic meaning to data without consuming large quantities of resources, and for a way to improve semantic understanding as information develops.

SUMMARY OF THE INVENTION

To find a context in which to answer a question, a directed set is constructed. The directed set comprises a plurality of elements and chains relating the concepts. One concept is identified as a maximal element. Chains are established in the directed set, connecting the maximal element to each concept in the directed set. More than one chain can connect the maximal element to each concept. A subset of the chains is selected to form a basis for the directed set. Each concept in the directed set is measured to determine how concretely each chain in the basis represents it. These measurements can be used to determine how closely related pairs of concepts are in the directed set.

The foregoing and other features, objects, and advantages of the invention will become more readily apparent from the following detailed description, which proceeds with reference to the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1A shows a computer system on which the invention can operate.

FIG. 1B shows the computer system of FIG. 1A connected to the Internet.

FIG. 2 shows the computer system of FIG. 1A listening to a content stream.

FIG. 3 shows an example of set of concepts that can form a directed set.

FIG. 4 shows a directed set constructed from the set of concepts of FIG. 3 in a preferred embodiment of the invention.

FIGs. 5A-5G show eight different chains in the directed set of FIG. 4 that form a basis for the directed set.

FIG. 6 is a flowchart of a method to construct a directed set in the system of FIG. 1A.

FIG. 7 is a flowchart of a method to add a new concept to a directed set in the system of FIG. 1A.

FIG. 8 is a flowchart of a method to update a basis for a directed set in the system of FIG. 1A.

FIG. 9 is a flowchart of a method of updating the concepts in a directed set in the system of FIG. 1A.

FIGs. 10A and 10B show how a new concept is added and relationships changed in the directed set of FIG. 4.

FIG. 11 is a flowchart of a method using a directed set in the system of FIG. 1A to help in answering a question.

FIG. 12 is a flowchart of a method using a directed set in the system of FIG. 1A to refine a query.

FIG. 13 shows data structures for storing a directed set, chains, and basis chains, such as the directed set of FIG. 3, the chains of FIG. 4, and the basis chains of FIGs. 5A-5G.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

FIG. 1A shows a computer system 105 on which a method and apparatus for using a multi-dimensional semantic space can operate. Computer system 105 conventionally includes a computer 110, a monitor 115, a keyboard 120, and a mouse 125. Optional equipment not shown in FIG. 1A can include a printer and other input/output devices. Also not shown in FIG. 1A are the conventional internal components of computer system 105: e.g., a central processing unit, memory, file system, etc.

Computer system 105 further includes a concept identification unit (CIU) 130, a chain unit (CU) 135, a basis unit (BU) 140, and a measurement unit (MU) 145. Concept identification unit 130 is responsible for identifying the concepts that will form a directed set, from which the multi-dimensional semantic space can be mapped. One concept is identified as a maximal element: this element describes (more or less concretely) every concept in the directed set. Chain unit 135 is responsible for constructing chains from the maximal element

to all other concepts identified by concept identification unit 130. Basis unit 140 is responsible for selecting a subset of the chains to form a basis for the directed set. Because basis unit 140 selects a subset of the chains established by chain unit 135, basis unit 140 is depicted as being part of chain unit 135. However, a person skilled in the art will recognize that basis unit 140 can be separate from chain unit 135. Measurement unit 145 is responsible for measuring how concretely each chain in the basis represents each concept. (How this measurement is performed is discussed below.) In the preferred embodiment, concept identification unit 130, chain unit 135, basis unit 140, and measurement unit 145 are implemented in software. However, a person skilled in the art will recognize that other implementations are possible. Finally, computer system 105 includes a data structure 150 (discussed with reference to FIG. 13 below). The data structure is responsible for storing the concepts, chains, and measurements of the directed set.

FIG. 1B shows computer system 105 connected over a network connection 140 to a network 145. The specifics of network connection 140 are not important, so long as the invention has access to a content stream to listen for concepts and their relationships. Similarly, computer system 105 does not have to be connected to a network 145, provided some content stream is available.

FIG. 2 shows computer system 105 listening to a content stream. In FIG. 2, network connection 140 includes a listening device 205. Listening device 205 (sometimes called a "listening mechanism") allows computer system 105 to listen to the content stream 210 (in FIG. 2, represented as passing through a "pipe" 215). Computer system 105 is parsing a number of concepts, such as "behavior," "female," "cat," "Venus Flytrap," "iguana," and so on. Listening device 205 also allows computer system 105 to determine the relationships between concepts.

But how is a computer, such as computer system 105 in FIGs. 1A, 1B, and 2 supposed to understand what the data it hears means? This is the question addressed below.

Semantic Value

Whether the data expressing content on the network is encoded as text, binary code, bit map or in any other form, there is a vocabulary that is either explicitly (such as for code) or implicitly (as for bitmaps) associated with the form. The vocabulary is more than an

arbitrarily-ordered list: an element of a vocabulary stands in relation to other elements, and the "place" of its standing is the *semantic value* of the element. For example, consider a spoon. Comparing the spoon with something taken from another scene – say, a shovel – one might classify the two items as being somewhat similar. And to the extent that form follows function in both nature and human artifice, this is correct! The results would be similar if the spoon were compared with a ladle. All three visual elements – the spoon, the shovel and the ladle – are *topologically* equivalent; each element can be transformed into the other two elements with relatively little geometric distortion.

What happens when the spoon is compared with a fork? Curiously enough, both the spoon and the fork are topologically equivalent. But comparing the ratio of boundary to surface area reveals a distinct contrast. In fact, the attribute (boundary)/(surface area) is a crude analog of the *fractal dimension* of the element boundary.

Iconic Representation

Fractal dimension possesses a nice linear ordering. For example, a space-filling boundary such as a convoluted coastline (or a fork!) would have a higher fractal dimension than, say, the boundary of a circle. Can the topology of an element be characterized in the same way? In fact, one *can* assign a topological measure to the vocabulary elements, but the measure may involve aspects of homotopy and homology that preclude a simple linear ordering. Suppose, for visual simplicity, that there is some simple, linearly ordered way of measuring the topological essence of an element. One can formally represent an *attribute space* for the elements, where *fork-like* and *spoon-like* resolve to different regions in the attribute space. In this case, one might adopt the standard Euclidean metric for $\mathbb{R}^2$ with one axis for "fractal dimension" and another for "topological measure," and thus have a well-defined notion of *distance* in attribute space. Of course, one must buy into all the hidden assumptions of the model. For example, is the orthogonality of the two attributes justified, i.e., are the attributes truly independent?

The example attribute space is a (simplistic) illustration of a semantic space, also known as a *concept space*. Above, the concern was with a vocabulary for human visual elements: a kind of *visual lexicon*. In fact, many researchers have argued for an *iconic representation* of meaning, particularly those looking for a representation unifying perception

and language. They take an empirical positivist position that *meaning* is simply an artifact of the "binding" of language to perception, and point out that all writing originated with pictographs (even the letter "A" is just an inverted ox head!). With the exception of some very specialized vocabularies, it is an unfortunate fact that most iconic models have fallen well short of the mark. What is the visual imagery for the word "maybe"? For that matter, the above example iconic model has shown how spoons and forks are different, but how does it show them to be the same (i.e., cutlery)?

Propositional Representation

Among computational linguists, a leading competitive theory to iconic representation is *propositional representation*. A proposition is typically framed as a pairing of an *argument* and a *predicate*. For example, the fragment "a red car" could be represented propositionally as the argument "a car" paired with the predicate "is red." The proposition simply asserts a property (the predicate) of an object (the argument). In this example, stipulating the argument alone has consequences; "a car" invokes the existential quantifier, and asserts instances for all relevant primitive attributes associated with the lexical element "car."

How about a phrase such as "every red car"? Taken by itself, the phrase asserts nothing – not even existence! It is a null proposition, and can be safely ignored. What about "every red car has a radio"? This is indeed making an assertion of sorts, but it is asserting a property of the semantic space itself; i.e., it is a meta-proposition. One can not instantiate a red car without a radio, nor can one remove a radio from a red car without either changing the color or losing the "car-ness" of the object. Propositions that are interpreted as assertions rather than as descriptions are called "meaning postulates."

At this point the reader should begin to suspect the preeminent role of the predicate, and indeed would be right to do so. Consider the phrase, "the boy hit the baseball."

nominative:	the boy	→ (is human), (is ~adult), (is male), (is ~infant), etc.
predicate:	(hit the baseball)	→
verb:	hit	→ (is contact), (is forceful), (is aggressive), etc.
d.o.:	the baseball	→ (is round), (is leather), (is stitched), etc.

The phrase has been transformed into two sets of attributes: the nominative attributes and two subsets of predicate attributes (verb and object). This suggests stipulating that all propositions must have the form (n: n ∈ N, p: p ∈ P), where N (the set of nominatives) is some appropriately restricted subset of $\wp$(P) (the power set of the space P of predicates). N is restricted to avoid things like ((is adult) and (is ~adult)). In this way the predicates can be used to generate a semantic space. A semantic representation might even be possible for something like, "The movie *The Boy Hit the Baseball* hit this critic's heart-strings!"

Given that propositions can be resolved to sets of predicates, the way forward becomes clearer. If one were to characterize sets of predicates as *clusters* of points in an attribute space along with some notion of *distance* between clusters, *one could quantify how close any two propositions are to each other*. This is the Holy Grail.

Before leaving this section, observe that another useful feature of the propositional model is *hierarchy of scope*, at least at the sentence level and below. Consider the phrase, "the boy hit the spinning baseball." The first-tier proposition is "x hit y." The second-tier propositions are "x is-a boy," and "y is-a baseball." The third-tier proposition is "y is spinning." By restricting the scope of the semantic space, attention can be focused on "hitting," "hitting spinning things," "people hitting things," etc.

Hyponymy & Meaning Postulates - Mechanisms for Abstraction

Two elements of the lexicon are related by *hyponymy* if the meaning of one is included in the meaning of the other. For example, the words "cat" and "animal" are related by hyponymy. A cat *is* an animal, and so "cat" is a hyponym of "animal."

A particular lexicon may not explicitly recognize some hyponymies. For example, the words "hit," "touch," "brush, " "stroke, " "strike," and "ram" are all hyponyms of the *concept* "co-incident in some space or context." Such a concept can be formulated as a *meaning postulate*, and the lexicon is extended with the meaning postulate in order to capture formally the hyponymy.

Note that the words "hit" and "strike" are also hyponyms of the word "realize" in the popular vernacular. Thus, lexical elements can surface in different hyponymies depending on the inclusion chain that is followed.

Topological Considerations

Now consider the *metrization problem*: how is the distance between two propositions determined? Many people begin by identifying a set S to work with (in this case, S = P, the set of predicates), and define a *topology* on S. A topology is a set O of subsets of S that satisfies the following criteria:

- Any union of elements of O is in O.
- Any finite intersection of elements of O is in O.
- S and the empty set are both in O.

The elements of O are called the *open* sets of S. If X is a subset of S, and p is an element of S, then p is called a *limit point* of X if every open set that contains p also contains a point in X distinct from p.

Another way to characterize a topology is to identify a *basis* for the topology. A set B of subsets of S is a basis if

- S = the union of all elements of B,
- for $p \in b_\alpha \cap b_\gamma$, ($b_\alpha$, $b_\gamma \in$ B), there exists $b_\lambda \in$ B such that $p \in b_\lambda$ and $b_\lambda \supseteq b_\alpha \cap b_\gamma$.

A subset of S is open if it is the union of elements of B. This defines a topology on S. Note that it is usually easier to characterize a basis for a topology rather than to explicitly identify all open sets. The space S is said to be *completely separable* if it has a countable basis.

It is entirely possible that there are two or more characterizations that yield the same topology. Likewise, one can choose two seemingly closely-related bases that yield nonequivalent topologies. As the keeper of the Holy Grail said to Indiana Jones, "Choose wisely!"

The goal is to choose as *strong* a topology as possible. Ideally, one looks for a compact metric space. One looks to satisfy separability conditions such that the space S is guaranteed to be *homeomorphic* to a subspace of Hilbert space (i.e., there is a continuous and one-to-one mapping from S to the subspace of Hilbert space). One can then adopt the Hilbert space metric. Failing this, as much structure as possible is imposed. To this end, consider the following axioms (the so-called "trennungaxioms").

- T_0. Given two points of a topological space S, at least one of them is contained in an open set not containing the other.

- T_1. Given two points of S, each of them lies in an open set not containing the other.
- T_2. Given two points of S, there are *disjoint* open sets, each containing just one of the two points (Hausdorff axiom).
- T_3. If C is a closed set in the space S, and if p is a point not in C, then there are disjoint open sets in S, one containing C and one containing p.
- T_4. If H and K are disjoint closed sets in the space S, then there are disjoint open sets in S, one containing H and one containing K.

Note that a set X in S is said to be *closed* if the complement of X is open. Since the intention is not to take the reader through the equivalent of a course in topology, simply observe that the distinctive attributes of T_3 and T_4 spaces are important enough to merit a place in the mathematical lexicon – T_3 spaces are called regular spaces, and T_4 spaces are called normal spaces – and the following very beautiful theorem:

- **Theorem 1**. Every completely separable regular space can be imbedded in a Hilbert coordinate space.

So, if there is a countable basis for S that satisfies T_3, then S is metrizable. The metrized spaced S is denoted as $(\mathbf{S}, d)$.

Finally, consider $\mathcal{H}(\mathbf{S})$, the set of all compact (non-empty) subsets of $(\mathbf{S}, d)$. Note that for $u, v \in \mathcal{H}(\mathbf{S})$, $u \cup v \in \mathcal{H}(\mathbf{S})$; i.e., the union of two compact sets is itself compact.

Define the pseudo-distance $\xi(x, u)$ between the point $x \in \mathbf{S}$ and the set $u \in \mathcal{H}(\mathbf{S})$ as

$$\xi(x, u) = \min\{d(x, y) : y \in u\}.$$

Using ξ define another pseudo-distance $\lambda(u, v)$ from the set $u \in \mathcal{H}(\mathbf{S})$ to the set $v \in \mathcal{H}(\mathbf{S})$:

$$\lambda(u, v) = \max\{\xi(x, v) : x \in u\}.$$

Note that in general it is *not* true that $\lambda(u, v) = \lambda(v, u)$. Finally, define the distance $h(u, v)$ between the two sets $u, v \in \mathcal{H}(\mathbf{S})$ as

$$h(u, v) = \max\{\lambda(u, v), \lambda(v, u)\}.$$

The distance function h is called the *Hausdorff* distance. Since

$h(u, v) = h(v, u)$,

$0 < h(u, v) < \infty$ for all $u, v \in \mathcal{H}(\mathbf{S})$, $u \neq v$,

$h(u, u) = 0$ for all $u \in \mathcal{H}(\mathbf{S})$,

$h(u, v) \leq h(u, w) + h(w, v)$ for all $u, v, w \in \mathcal{H}(\mathbf{S})$,

the metric space $(\mathcal{H}(\mathbf{S}), h)$ can now be formed. The completeness of the underlying metric space $(\mathbf{S}, d)$ is sufficient to show that every Cauchy sequence $\{u_k\}$ in $(\mathcal{H}(\mathbf{S}), h)$ converges to a point in $(\mathcal{H}(\mathbf{S}),h)$. Thus, $(\mathcal{H}(\mathbf{S}), h)$ is a *complete* metric space.

If S is metrizable, then it is $(\mathcal{H}(\mathbf{S}), h)$ wherein lurks that elusive beast, *semantic value*. For, consider the two propositions, $\rho_1 = (n_1, p_1)$, $\rho_2 = (n_2, p_2)$. Then the *nominative distance* $|n_2 - n_1|$ can be defined as $h(\overline{n_1}, \overline{n_2})$, where $\overline{n}$ denotes the closure of n. The *predicate distance* can be defined similarly. Finally, one might define:

$$|\rho_2 - \rho_1| = (|n_2 - n_1|^2 + |p_2 - p_1|^2)^{1/2} \qquad \text{Equation (1a)}$$

or alternatively one might use "city block" distance:

$$|\rho_2 - \rho_1| = |n_2 - n_1| + |p_2 - p_1| \qquad \text{Equation (1b)}$$

as a fair approximation of distance. Those skilled in the art will recognize that other metrics are also possible: for example:

$$(\Sigma(\rho_{2,i} - \rho_{1,i})^{n})^{1/n} \qquad \text{Equation (1c)}$$

The reader may recognize $(\mathcal{H}(\mathbf{S}), h)$ as the *space of fractals*. Some compelling questions come immediately to mind. Might one be able to find *submonoids* of *contraction mappings* corresponding to *related* sets in $(\mathcal{H}(\mathbf{S}), h)$; related, for example, in the sense of convergence to the same collection of *attractors*? This could be a rich field to plow.

An Example Topology

Consider an actual topology on the set P of predicates. This is accomplished by exploiting the notion of hyponymy and meaning postulates.

Let P be the set of predicates, and let B be the set of all elements of 2^{2^P}, i.e., $\wp(\wp(P))$, that express hyponymy. B is a basis, if not of 2^P, i.e., $\wp(P)$, then at least of everything worth talking about: $S = \cup (b: b \in B)$. If b_α, $b_\gamma \in B$, neither containing the other, have a non-empty intersection that is not already an explicit hyponym, extend the basis B with the meaning postulate $b_\alpha \cap b_\gamma$. For example, "dog" is contained in both "carnivore" and "mammal." So, even though the core lexicon may not include an entry equivalent to "carnivorous mammal," it is a worthy meaning postulate, and the lexicon can be extended to include the intersection. Thus, B is a basis for S.

Because hyponymy is based on nested subsets, there is a hint of partial ordering on S. A partial order would be a big step towards establishing a metric.

At this point, a concrete example of a (very restricted) lexicon is in order. FIG. 3 shows a set of concepts, including "thing" 305, "man" 310, "girl" 312, "adult human" 315, "kinetic energy" 320, and "local action" 325. "Thing" 305 is the maximal element of the set, as every other concept is a type of "thing." Some concepts, such as "man" 310 and "girl" 312 are "leaf concepts," in the sense that no other concept in the set is a type of "man" or "girl." Other concepts, such as "adult human" 315, "kinetic energy" 320, and "local action" 325 are "internal concepts," in the sense that they are types of other concepts (e.g., "local action" 325 is a type of "kinetic energy" 320) but there are other concepts that are types of these concepts (e.g., "man" 310 is a type of "adult human" 315).

FIG. 4 shows a directed set constructed from the concepts of FIG. 3. For each concept in the directed set, there is at least one chain extending from maximal element "thing" 305 to the concept. These chains are composed of directed links, such as links 405, 410, and 415, between pairs of concepts. In the directed set of FIG. 4, every chain from maximal element "thing" must pass through either "energy" 420 or "category" 425. Further, there can be more than one chain extending from maximal element "thing" 305 to any concept. For example, there are four chains extending from "thing" 305 to "adult human"

315: two go along link 410 extending out of "being" 435, and two go along link 415 extending out of "adult" 445.

Some observations about the nature of FIG. 4:

- First, the model is a *topological space*.
- Second, note that *the model is not a tree*. In fact, it is an example of a *directed set*. For example, concepts "being" 430 and "adult human" 315 are types of multiple concepts higher in the hierarchy. "Being" 430 is a type of "matter" 435 and a type of "behavior" 440; "adult human" 315 is a type of "adult" 445 and a type of "human" 450.
- Third, observe that the relationships expressed by the links are indeed relations of hyponymy.
- Fourth, note particularly – but without any loss of generality – that "man" 310 maps to both "energy" 420 and "category" 425 (via composite mappings) which in turn both map to "thing" 305; i.e., the (composite) relations are multiple valued and induce a partial ordering. These multiple mappings are natural to the meaning of things and critical to semantic characterization.
- Finally, note that "thing" 305 is *maximal*; indeed, "thing" 305 is the *greatest* element of *any* quantization of the lexical semantic field (subject to the premises of the model).

Metrizing **S**

FIGs. 5A-5G show eight different chains in the directed set that form a basis for the directed set. FIG. 5A shows chain 505, which extends to concept "man" 310 through concept "energy" 420. FIG. 5B shows chain 510 extending to concept "iguana." FIG. 5C shows another chain 515 extending to concept "man" 310 via a different path. FIGs. 5D-5G show other chains.

FIG. 13 shows a data structure for storing the directed set of FIG. 3, the chains of FIG. 4, and the basis chains of FIGs. 5A-5G. In FIG. 13, concepts array 1305 is used to store the concepts in the directed set. Concepts array 1305 stores pairs of elements. One element identifies concepts by name; the other element stores numerical identifiers 1306. For example, concept name 1307 stores the concept "dust," which is paired with numerical

identifier "2" 1308. Concepts array 1305 shows 9 pairs of elements, but there is no theoretical limit to the number of concepts in concepts array 1305. In concepts array 1305, there should be no duplicated numerical identifiers 1306. In FIG. 13, concepts array 1305 is shown sorted by numerical identifier 1306, although this is not required. When concepts array 1305 is sorted by numerical identifier 1306, numerical identifier 1306 can be called the *index* of the concept name.

Maximal element (ME) 1310 stores the index to the maximal element in the directed set. In FIG. 13, the concept index to maximal element 1310 is "6," which corresponds to concept "thing," the maximal element of the directed set of FIG. 4.

Chains array 1315 is used to store the chains of the directed set. Chains array 1315 stores pairs of elements. One element identifies the concepts in a chain by index; the other element stores a numerical identifier. For example, chain 1317 stores a chain of concept indices "6", "5", "9", "7", and "2," and is indexed by chain index "1" (1318). (Concept index 0, which does not occur in concepts array 1305, can be used in chains array 1315 to indicate the end of the chain. Additionally, although chain 1317 includes five concepts, the number of concepts in each chain can vary.) Using the indices of concepts array 1305, this chain corresponds to concepts "thing," "energy," "potential energy," "matter," and "dust." Chains array 1315 shows one complete chain and part of a second chain, but there is no theoretical limit to the number of chains stored in chain array 1315. Observe that, because maximal element 1310 stores the concept index "6," every chain in chains array 1315 should begin with concept index "6." Ordering the concepts within a chain is ultimately helpful in measuring distances between the concepts. However concept order is not required. Further, there is no required order to the chains as they are stored in chains array 1315.

Basis chains array 1320 is used to store the chains of chains array 1315 that form a basis of the directed set. Basis chains array 1320 stores chain indices into chains array 1315. Basis chains array 1320 shows four chains in the basis (chains 1, 4, 8, and 5), but there is no theoretical limit to the number of chains in the basis for the directed set.

Euclidean distance matrix 1325A stores the distances between pairs of concepts in the directed set of FIG. 4. (How distance is measured between pairs of concepts in the directed set is discussed below. But in short, the concepts in the directed set are mapped to state vectors in multi-dimensional space, where a state vector is a directed line segment starting at

the origin of the multi-dimensional space and extending to a point in the multi-dimensional space.) The distance between the end points of pairs of state vectors representing concepts is measured. The smaller the distance is between the state vectors representing the concepts, the more closely related the concepts are. Euclidean distance matrix 1325A uses the indices 1306 of the concepts array for the row and column indices of the matrix. For a given pair of row and column indices into Euclidean distance matrix 1325A, the entry at the intersection of that row and column in Euclidean distance matrix 1325A shows the distance between the concepts with the row and column concept indices, respectively. So, for example, the distance between concepts "man" and "dust" can be found at the intersection of row 1 and column 2 of Euclidean distance matrix 1325A as approximately 1.96 units. The distance between concepts "man" and "iguana" is approximately 1.67, which suggests that "man" is closer to "iguana" than "man" is to "dust." Observe that Euclidean distance matrix 1325A is symmetrical: that is, for an entry in Euclidean distance matrix 1325A with given row and column indices, the row and column indices can be swapped, and Euclidean distance matrix 1325A will yield the same value. In words, this means that the distance between two concepts is not dependent on concept order: the distance from concept "man" to concept "dust" is the same as the distance from concept "dust" to concept "man."

Angle subtended matrix 1325B is an alternative way to store the distance between pairs of concepts. Instead of measuring the distance between the state vectors representing the concepts (see below), the angle between the state vectors representing the concepts is measured. This angle will vary between 0 and 90 degrees. The narrower the angle is between the state vectors representing the concepts, the more closely related the concepts are. As with Euclidean distance matrix 1325A, angle subtended matrix 1325B uses the indices 1306 of the concepts array for the row and column indices of the matrix. For a given pair of row and column indices into angle subtended matrix 1325B, the entry at the intersection of that row and column in angle subtended matrix 1325B shows the angle subtended the state vectors for the concepts with the row and column concept indices, respectively. For example, the angle between concepts "man" and "dust" is approximately 51 degrees, whereas the angle between concepts "man" and "iguana" is approximately 42 degrees. This suggests that "man" is closer to "iguana" than "man" is to "dust." As with Euclidean distance matrix 1325A, angle subtended matrix 1325B is symmetrical.

Not shown in FIG. 13 is a data structure component for storing state vectors (discussed below). As state vectors are used in calculating the distances between pairs of concepts, if the directed set is static (i.e., concepts are not being added or removed and basis chains remain unchanged), the state vectors are not required after distances are calculated. Retaining the state vectors is useful, however, when the directed set is dynamic. A person skilled in the art will recognize how to add state vectors to the data structure of FIG. 13.

Although the data structure for concepts array 1305, maximal element 1310 chains array 1315, and basis chains array 1320 in FIG. 13 are shown as arrays, a person skilled in the art will recognize that other data structures are possible. For example, concepts array could store the concepts in a linked list, maximal element 1310 could use a pointer to point to the maximal element in concepts array 1305, chains array 1315 could use pointers to point to the elements in concepts array, and basis chains array 1320 could use pointers to point to chains in chains array 1315. Also, a person skilled in the art will recognize that the data in Euclidean distance matrix 1325A and angle subtended matrix 1325B can be stored using other data structures. For example, a symmetric matrix can be represented using only one half the space of a full matrix if only the entries below the main diagonal are preserved and the row index is always larger than the column index. Further space can be saved by computing the values of Euclidean distance matrix 1325A and angle subtended matrix 1325B "on the fly" as distances and angles are needed.

Returning to FIGs. 5A-5G, how are distances and angles subtended measured? The chains shown in FIGs. 5A-5G suggest that the relation between any node of the model and the maximal element "thing" 305 can be expressed as any one of a set of *composite* functions; one function for each chain from the minimal node μ to "thing" 305 (the n^{th} predecessor of μ along the chain):

$$f: \mu \Rightarrow thing = f_1 \circ f_2 \circ f_3 \circ \cdots \circ f_n$$

where the chain connects $n + 1$ concepts, and f_j: links the $(n - j)^{th}$ predecessor of μ with the $(n + 1 - j)^{th}$ predecessor of μ, $1 \leq j \leq n$. For example, with reference to FIG. 5A, chain 505 connects nine concepts. For chain 505, f_1 is link 505A, f_2 is link 505B, and so on through f_8 being link 505H.

Consider the set of all such functions for all minimal nodes. Choose a countable subset $\{f_k\}$ of functions from the set. For each f_k construct a function g_k: $S \Rightarrow \mathbf{I}^1$ as follows.

For $s \in S$, s is in relation (under hyponymy) to "thing" 305. Therefore, s is in relation to at least one predecessor of μ, the minimal element of the (unique) chain associated with f_k. Then there is a predecessor of smallest index (of μ), say the m^{th}, that is in relation to s. Define:

$$g_k(s) = (n - m) / n \qquad \text{Equation (2)}$$

This formula gives a measure of concreteness of a concept to a given chain associated with function f_k.

As an example of the definition of g_k, consider chain 505 of FIG. 5A, for which n is 8. Consider the concept "cat" 555. The smallest predecessor of "man" 310 that is in relation to "cat" 555 is "being" 430. Since "being" 430 is the fourth predecessor of "man" 310, m is 4, and g_k("cat" 555) = (8 - 4) / 8 = ½. "Iguana" 560 and "plant" 560 similarly have g_k values of ½. But the only predecessor of "man" 310 that is in relation to "adult" 445 is "thing" 305 (which is the eighth predecessor of "man" 310), so m is 8, and g_k("adult" 445) = 0.

Finally, define the vector valued function $\varphi: S \Rightarrow \mathbb{R}^k$ relative to the indexed set of scalar functions $\{g_1, g_2, g_3, \ldots, g_k\}$ (where scalar functions $\{g_1, g_2, g_3, \ldots, g_k\}$ are defined according to Equation (2)) as follows:

$$\varphi(s) = \langle g_1(s), g_2(s), g_3(s), \ldots, g_k(s) \rangle \qquad \text{Equation (3)}$$

This state vector $\varphi(s)$ maps a concept s in the directed set to a point in k-space ($\mathbb{R}^k$). One can measure distances between the points (the state vectors) in k-space. These distances provide measures of the closeness of concepts within the directed set. The means by which distance can be measured include distance functions, such as Equations (1a), (1b), or (1c). Further, trigonometry dictates that the distance between two vectors is related to the angle subtended between the two vectors, so means that measure the angle between the state vectors also approximates the distance between the state vectors. Finally, since only the direction (and not the magnitude) of the state vectors is important, the state vectors can be normalized to the unit sphere. If the state vectors are normalized, then the angle between two state vectors is no longer an approximation of the distance between the two state vectors, but rather is an exact measure.

The functions g_k are analogous to step functions, and in the limit (of refinements of the topology) the functions are continuous. Continuous functions preserve local topology; i.e., "close things" in S map to "close things" in $\mathbb{R}^k$, and "far things" in S tend to map to "far things" in $\mathbb{R}^k$.

Example Results

The following example results show state vectors $\varphi(s)$ using chain 505 as function g_1, chain 510 as function g_2, and so on through chain 540 as function g_8.

φ("boy") $\Rightarrow$ $\langle 3/4, 5/7, 4/5, 3/4, 7/9, 5/6, 1, 6/7\rangle$

φ("dust") $\Rightarrow$ $\langle 3/8, 3/7, 3/10, 1, 1/9, 0, 0, 0\rangle$

φ("iguana") $\Rightarrow$ $\langle 1/2, 1, 1/2, 3/4, 5/9, 0, 0, 0\rangle$

φ("woman") $\Rightarrow$ $\langle 7/8, 5/7, 9/10, 3/4, 8/9, 2/3, 5/7, 5/7\rangle$

φ("man") $\Rightarrow$ $\langle 1, 5/7, 1, 3/4, 1, 1, 5/7, 5/7\rangle$

Using these state vectors, the distances between concepts and the angles subtended between the state vectors are as follows:

Pairs of Concepts	Distance (Euclidean)	Angle Subtended
"boy" and "dust"	~1.85	~52°
"boy" and "iguana"	~1.65	~46°
"boy" and "woman"	~0.41	~10°
"dust" and "iguana"	~0.80	~30°
"dust" and "woman"	~1.68	~48°
"iguana" and "woman"	~1.40	~39°
"man" and "woman"	~0.39	~07°

From these results, the following comparisons can be seen:

- "boy" is closer to "iguana" than to "dust."
- "boy" is closer to "iguana" than "woman" is to "dust."
- "boy" is much closer to "woman" than to "iguana" or "dust."
- "dust" is further from "iguana" than "boy" to "woman" or "man" to "woman."
- "woman" is closer to "iguana" than to "dust."

- "woman" is closer to "iguana" than "boy" is to "dust."
- "man" is closer to "woman" than "boy" is to "woman."

All other tests done to date yield similar results. The technique works consistently well.

How It (Really) Works

As described above, construction of the φ transform is (very nearly) an algorithm. In effect, this describes a *recipe* for metrizing a lexicon – or for that matter, metrizing anything that can be modeled as a directed set – but does not address the issue of *why* it works. In other words, *what's really going on here*? To answer this question, one must look to the underlying mathematical principles.

First of all, what is the nature of **S**? Earlier, it was suggested that a propositional model of the lexicon has found favor with many linguists. For example, the lexical element "automobile" might be modeled as:

{automobile: *is a* machine,
is a vehicle,
has engine,
has brakes,
...
}

In principle, there might be infinitely many such properties, though practically speaking one might restrict the cardinality to $\aleph_0$ (countably infinite) in order to ensure that the properties are addressable. If one were disposed to do so, one might require that there be only finitely many properties associated with a lexical element. However, there is no compelling reason to require finiteness.

At any rate, one can see that "automobile" is simply an element of the power set of **P**, the set of all propositions; i.e., it is an element of the set of all subsets of **P**. The power set is denoted as $\wp(\mathbf{P})$. Note that the first two properties of the "automobile" example express "*is a*" relationships. By "*is a*" is meant entailment. *Entailment* means that, were one to intersect the properties of every element of $\wp(\mathbf{P})$ that is called, for example, "machine," then the

intersection would contain a subset of properties common to anything (in $\wp(\mathbf{P})$) that one has, does, will or would have called "machine." Reliance on the existence of a "least" common subset of properties to define entailment has a hint of *well ordering* about it; and indeed it is true that the *axiom of choice* is relied on to define entailment.

For the moment, restrict the notion of *meaning postulate* to that of entailment. Let $\mathbf{B} = \{b_\alpha\}$ be the set of elements of $\wp(\wp(\mathbf{P}))$ that correspond to *good* meaning postulates; e.g., $b_m \in \mathbf{B}$ is the set of all elements of $\wp(\mathbf{P})$ that entail "machine." By "good" is meant *complete* and *consistent*. "Complete" means non-exclusion of objects that *should* entail (some concept). "Consistent" means exclusion of objects that *should not* entail (any concept). *Should/should-not* are understood to be negotiated between the community (of language users) and its individuals.

Note that if the intersection of b_β and b_γ is non-empty, then $b_\beta \cap b_\gamma$ is a "good" meaning postulate, and so must be in **B**. Define the set $\mathbf{S} = \cup\, b_\alpha$ to be the lexicon. A point of **S** is an element of $\wp(\mathbf{P})$ that entails at least one meaning postulate.

B was deliberately constructed to be the basis of a topology τ for **S**. In other words, an open set in **S** is defined to be the union of elements of **B**. This is what is meant when one says that hyponymy is used to define the topology of the lexicon (in this particular embodiment).

The separability properties of **S** are reflected in the Genus/Species relationships of the unfolding inclusion chains. The T_0-T_4 trennungsaxioms are adopted. Now consider the set of bounded continuous real valued functions on **S**.

- **Urysohn's lemma**. If **S** is a normal space and A and B are two disjoint closed subsets of **S**, then there is a real-valued continuous function g: $\mathbf{S} \Rightarrow \mathbf{I}^1$ of **S** into the unit interval $\mathbf{I}^1$ such that $g(A) = 0$ and $g(B) = 1$.

The use of g to denote the function was not accidental; it should evoke the scalar coordinate functions $\{g_1, g_2, g_3, \ldots, g_k\}$ defined per Equation (2) above. A proof of the lemma can be found in almost any elementary general topology book.

The end is in sight! Before invoking a final theorem of Urysohn's and completing the metrization of **S**, the notion of a Hilbert coordinate space must be introduced.

Consider the set **H** of all sequences $\gamma = \{\gamma_1, \gamma_2, \gamma_3, \ldots\}$ such that $\Sigma^{\infty} \gamma_i^2$ converges. Define the metric:

$$d(\gamma, \chi) = (\sum_{i=1}^{\infty} (\gamma_i - \chi_i)^2)^{1/2}$$

on the set **H**, and denote the Hilbert coordinate space (**H**, d).

If the sequence $\{\gamma_1, \gamma_2, \gamma_3, \ldots\}$ is considered as a vector, one can think of Hilbert space as a kind of "super" Euclidean space. Defining vector addition and scalar multiplication in the usual way, it is no great feat to show that the resultant vector is in **H**. Note that the standard inner product works just fine.

Before the metric space equivalent to the topological space (**S**, τ) can be found, one last theorem is needed.

- **Theorem 2**. A T_1-space S is regular if and only if for each point p in S and each open set U containing p, there is an open set V containing p whose closure $\overline{V}$ is contained in U.

In looking for a metric space equivalent to the topological space (**S**, τ), Urysohn's lemma should be a strong hint to the reader that perhaps (**H**, d) should be considered.

- **Theorem 3**. Every completely separable normal space **S** is homeomorphic to a subspace of Hilbert's coordinate space.

This theorem is proven by actually constructing the homeomorphism.

Proof: Let $B_1, B_2, \ldots, B_n, \ldots$ be a countable basis for **S**. In view of Theorem 2, there are pairs B_i, B_j, such that $\overline{B_i}$ is contained in B_j; in fact, each point of point of **S** lies in infinitely many such pairs, or is itself an open set. However, there are at most a countable number of pairs for each point of **S**. For each such pair B_i and B_j, Urysohn's lemma provides a function g_n of **S** into $\mathbf{I}^1$ with the property that $g_n(\overline{B_i}) = 0$ and $g_n(S - B_j) = 1$. (If the point p forms an open set, then take $g_n = 0$ for large n.) Letting **H** denote the Hilbert coordinate space, define the (vector-valued) mapping ϑ of **S** into **H** by setting

$$\vartheta(s) = \{g_1(s), g_2(s)/2, g_3(s)/3, \ldots, g_n(s)/n, \ldots\}$$

for each point s in **S**. It remains to prove that the function ϑ so defined is continuous, one-to-one, and open.

The original proof (in its entirety) of Theorem 3 is available in the literature. When ϑ is applied to a lexicon with the entailment topology, it is herein called the Bohm transformation. Clearly, the finite-dimensional transform φ is an approximation of the Bohm transform, mapping the explicate order of the lexicon to a (shallow) implicate order in $\mathbb{R}^k$.

Now that the mathematical basis for constructing and using a lexicon has been presented, the process of constructing the lexical semantic space can be explained. FIG. 6 is a flowchart of the steps to construct a directed set. At step 605, the concepts that will form the basis for the semantic space are identified. These concepts can be determined according to a heuristic, or can be defined statically. At step 610, one concept is selected as the maximal element. At step 615, chains are established from the maximal element to each concept in the directed set. As noted earlier, there can be more than one chain from the maximal element to a concept: the directed set does not have to be a tree. Also, as discussed above, the chains represent a topology that allows the application of Uryshon's lemma to metrize the set: for example, hyponomy, meronomy, or any other relations that induce inclusion chains on the set. At step 620, a subset of the chains is selected to form a basis for the directed set. At step 625, each concept is measured to see how concretely each basis chain represents the concept. Finally, at step 630, a state vector is constructed for each concept, where the state vector includes as its coordinates the measurements of how concretely each basis chain represents the concept.

FIG. 7 is a flowchart of how to add a new concept to an existing directed set. At step 705, the new concept is added to the directed set. The new concept can be learned by any number of different means. For example, the administrator of the directed set can define the new concept. Alternatively, the new concept can be learned by listening to a content stream as shown in FIG. 2. A person skilled in the art will recognize that the new concept can be learned in other ways as well. The new concept can be a "leaf concept" or an "intermediate concept." Recall that an "intermediate concept" is one that is an abstraction of further concepts; a "leaf concept" is one that is not an abstraction of further concepts. For example, referring to FIG. 4, "man" 310 is a "leaf concept," but "adult human" 315 is an "intermediate concept. Returning to FIG. 7, at step 710, a chain is established from the maximal element to the new concept. Determining the appropriate chain to establish to the new concept can be

done manually or based on properties of the new concept learned by the system. A person skilled in the art will also recognize that, as discussed above, more than one chain to the new concept can be established. At step 715, the new concept is measured to see how concretely each chain in the basis represents the new concept. Finally, at step 720, a state vector is created for the new concept, where the state vector includes as its coordinates the measurements of how concretely each basis chain represents the new concept.

FIG. 8 is a flowchart of how to update the basis, either by adding to or removing from the basis chains. If chains are to be removed from the basis, then at step 805 the chains to be removed are deleted. Otherwise, at step 810 new chains are added to the basis. If a new chain is added to the basis, each concept must be measured to see how concretely the new basis chain represents the concept (step 815). Finally, whether chains are being added to or removed from the basis, at step 820 the state vectors for each concept in the directed set are updated to reflect the change.

A person skilled in the art will recognize that, although FIG. 8 shows adding and removing basis chains to be separate operations, they can be done at the same time. In other words, one basis chain can be deleted and a new basis chain added at the same time.

FIG. 9 is a flowchart of how the directed set is updated. At step 905, the system is listening to a content stream. At step 910, the system parses the content stream into concepts. At step 915, the system identifies relationships between concepts in the directed set that are described by the content stream. Then, if the relationship identified at step 915 indicates that an existing chain is incorrect, at step 920 the existing chain is broken. Alternatively, if the relationship identified at step 915 indicates that a new chain is needed, at step 925 a new chain is established.

A person skilled in the art will recognize that, although FIG. 9 shows establishing new chains and breaking existing chains to be separate operations, they can be done at the same time. In other words, an identified relationship may require breaking an existing chain and establishing a new chain at the same time.

FIGs. 10A and 10B show how new concepts are added and relationships changed in the directed set of FIG. 4. FIGs. 10A and 10B show a close-up of a portion of the directed set of FIG. 4. FIG. 10A shows the state of the directed set after the system listens to the content stream 210 of FIG. 2. The terms "behavior," "female," "cat," "Venus Flytrap," and "iguana,"

are parsed from the content stream. For example, the stream may have included the question "How does the behavior of a female cat around a Venus Flytrap differ from that around an iguana?", from which the concepts were parsed. The term "Venus Flytrap" is unknown in the directed set, and a new concept "Venus Flytrap" 1005 is added to the directed set. The directed set may then conclude that, since "Venus Flytrap" is being compared to an "iguana," that "Venus Flytrap" is some type of animal, and should be related to "animal" 1010. (The directed set might even be more specific and conclude that "Venus Flytrap" is the same type of animal as "iguana," i.e., a reptile, but for this example a more general conclusion is assumed.) The directed set then introduces a chain 1015 through "animal" 1010 to "Venus Flytrap" 1005.

Assume that at this point, the directed set learns that a Venus Flytrap is some kind of plant, and not an animal. As shown in FIG. 10B, the directed set needs to establish a relationship between "Venus Flytrap" 1005 and "plant" 1020, and break the relationship with "animal" 1010. The directed set then breaks chain 1015 and adds chain 1025.

FIG. 11 shows a flowchart of how a directed set can be used to help in answering a question. At step 1105, the system receives the question. At step 1110, the system parses the question into concepts. At step 1115, the distances between the parsed concepts are measured in a directed set. Finally, at step 1120, using the distances between the parsed concepts, a context is established in which to answer the question.

FIG. 12 shows a flowchart of how a directed set can be used to refine a query, for example, to a database. At step 1205, the system receives the query. At step 1210, the system parses the query into concepts. At step 1215, the distances between the parsed concepts are measured in a directed set. At step 1220, using the distances between the parsed concepts, a context is established in which to refine the query. At step 1225, the query is refined according to the context. Finally, at step 1230, the refined query is submitted to the query engine.

Having illustrated and described the principles of our invention in a preferred embodiment thereof, it should be readily apparent to those skilled in the art that the invention can be modified in arrangement and detail without departing from such principles. We claim all modifications coming within the spirit and scope of the accompanying claims.

We claim:

1. A method for building a directed set to allow a user of a computer system to find a context in which to answer a question, the method comprising:

identifying a plurality of concepts to form a directed set, wherein one concept is a maximal element;

establishing chains in the directed set from the maximal element to each concept;

selecting one or more chains in the directed set as a basis; and

measuring how concretely each concept is represented in each chain in the basis.

2. A method according to claim 1 further comprising creating a state vector for each concept in the directed set, wherein each state vector includes as its components measures of how concretely the concept is represented in each chain in the basis.

3. A method according to claim 2 wherein creating a state vector for each concept in the directed set includes measuring a distance between the state vectors for each pair of concepts.

4. A method according to claim 1 further comprising introducing a new concept into the directed set.

5. A method according to claim 4 wherein introducing a new concept includes:

adding a new chain from the maximal element to the new concept; and

measuring new distances from the new concept to each chain in the basis.

6. A method according to claim 1 further comprising:

discarding the chains in the basis; and

re-selecting one or more chains in the directed set as a new basis.

7. A method according to claim 1 further comprising:

receiving new information about a first concept in the directed set; and

updating the directed links for the first concept.

8. A method according to claim 7 wherein updating the directed links includes at least one of:

a) removing an existing chain from the maximal element to the first concept; and

b) establishing a new chain from the maximal element to the first concept.

9. A method according to claim 1 wherein identifying a plurality of concepts includes:

listening to a content stream; and

parsing the concepts from the content stream.

10. A method according to claim 1 wherein establishing directed links between a first concept and a second concept includes:

listening to a content stream;

identifying a relationship between the first concept and the second concept from the content stream; and

establishing a chain from the maximal element to the first concept through the second concept to model the relationship between the first and second concepts.

11. A computer-readable medium containing a program to build a directed set to allow a user of a computer system to find a context in which to answer a question, the program comprising:

identification software to identify a plurality of concepts to form a directed set, wherein one concept is a maximal element;

chain-establishment software to establish chains in the directed set from the maximal element to each concept;

chain-selection software to select one or more chains in the directed set as a basis; and

measurement software to measure how concretely each concept is represented in each chain in the basis.

12. A storage medium for storing a lexicon as a directed set for use by an application program to establish a context for a query, the storage medium comprising:

a data structure stored in the storage medium, the data structure including the lexicon and including:

a plurality of concepts stored in the storage medium, wherein one concept is a maximal element; and

at least one chain extending from the maximal element to each concept, wherein the chain includes an ordered subset of the concepts, beginning with the maximal element and ending with the concept.

13. A storage medium according to claim 12 further comprising a data structure storing a plurality of distances between pairs of concepts.

14. A storage medium according to claim 12 further comprising a data structure identifying at least one chain as a chain in a basis for the directed set.

15. An apparatus on a computer system to build a directed set to allow a user of the computer system to find a context in which to answer a question, the apparatus comprising:

a data structure according to claim 12 to store the directed set;

an identification unit to identify the plurality of concepts in the directed set, wherein the directed set includes a maximal element;

a chain unit to establish chains in the directed set from the maximal element to each concept;

a basis unit to select one or more chains in the directed set as a basis; and

a measurement unit to measure how concretely each concept is represented in each chain in the basis.

16. An apparatus on a computer system to build a directed set to allow a user of the computer system to determine what questions can be answered using a given context, the apparatus comprising:

a data structure according to claim 12 to store the directed set;

an identification unit to identify the plurality of concepts in the directed set, wherein the directed set includes a maximal element;

a chain unit to establish chains in the directed set from the maximal element to each concept;

a basis unit to select one or more chains in the directed set as a basis; and

a measurement unit to measure how concretely each concept is represented in each chain in the basis.

17. A method for a user of a computer system to find a context to aid in answering a question, the method comprising:

parsing the question into one or more parsed concepts;

measuring distances in a directed set between the one or more parsed concepts; and

using the distances between the one or more parsed concepts to establish a context for the question.

18. A method according to claim 17 in which measuring distances in a directed set includes:

establishing one or more chains in the directed set, wherein each chain is rooted at a maximal element in the directed set and extends to a concept in the directed set;

creating a distance vector for the one or more parsed concepts in the directed set, wherein each distance vector includes as its components the measure of how concretely the concept is represented in each chain; and

measuring a distance between the distance vectors for each pair of parsed concepts.

19. A method for using a lexicon to submit a refined query input by a user to a query engine, wherein the refined query the method comprising:

parsing the query into one or more parsed concepts;

measuring distances in the lexicon between the one or more parsed concepts;

using the distances between the one or more parsed concepts to establish a context for the query;

refining the query according to the context for the query; and

submitting the refined query to the query engine.

20. An apparatus on a computer system to enable a user of the computer system to find a context in which to answer a question, the apparatus comprising:

a directed set stored in the computer system, the directed set including a plurality of first concepts, a maximal element, and at least one basis chain extending from the maximal element to one of the first concepts;

an input for receiving a content stream;

a listening mechanism listening to the content stream and parsing the content stream into second concepts; and

a measurement mechanism measuring distances between pairs of the second concepts according to the plurality of first concepts and the basis chains of the directed set.

21. An apparatus according to claim 20, wherein:

the apparatus further comprises a network connection; and

the input for receiving the content stream is coupled to the network connection.

22. An apparatus according to claim 20, wherein the measurement mechanism includes:

a state vector constructor converting each second concept into a state vector in Euclidean k-space; and

measuring means for measuring the distance between state vectors corresponding to the second concepts according to the plurality of first concepts and the basis chains of the directed set.

CONSTRUCTION, MANIPULATION, AND COMPARISON OF A MULTI-DIMENSIONAL SEMANTIC SPACE

ABSTRACT

A directed set can be used to establish contexts for linguistic concepts: for example, to aid in answering a question, to refine a query, or even to determine what questions can be answered given certain knowledge. A directed set includes a plurality of elements and chains relating the concepts. One concept is identified as a maximal element. The chains connect the maximal element to each concept in the directed set, and more than one chain can connect the maximal element to any individual concept either directly or through one or more intermediate concepts. A subset of the chains is selected to form a basis for the directed set. Each concept in the directed set is measured to determine how concretely each chain in the basis represents it. These measurements for a single concept form a vector in Euclidean k-space. Distances between these vectors can be used to determine how closely related pairs of concepts are in the directed set.

05220 e5ce1500

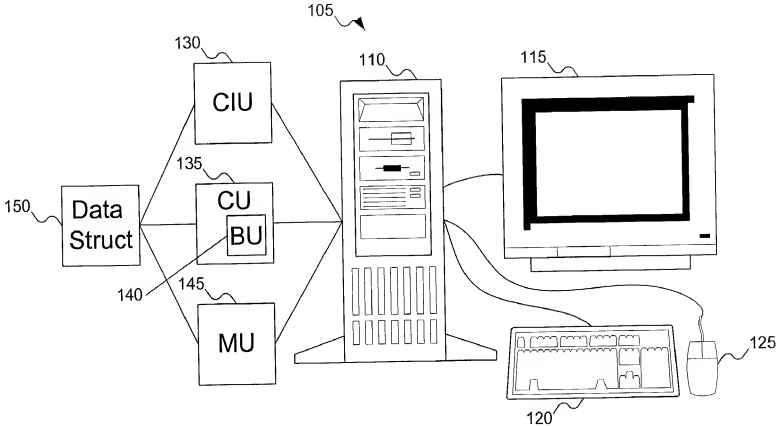


FIG. 1A

0052e0 e30e1500

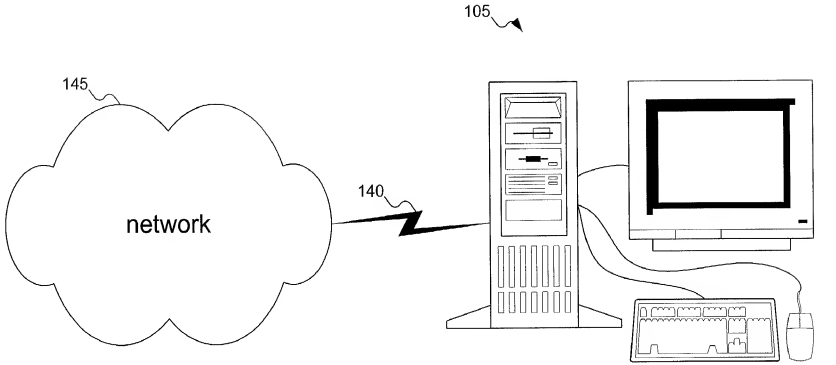


FIG. 1B

0052201-050705

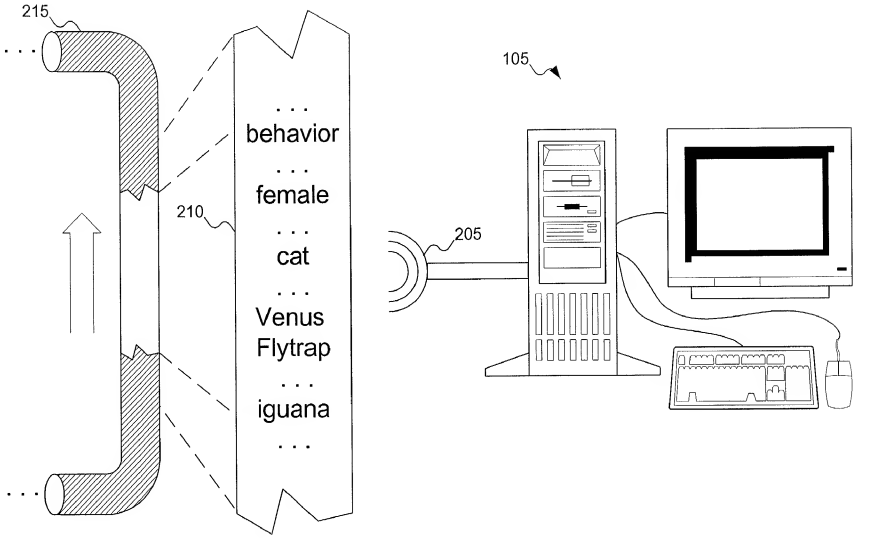


FIG. 2

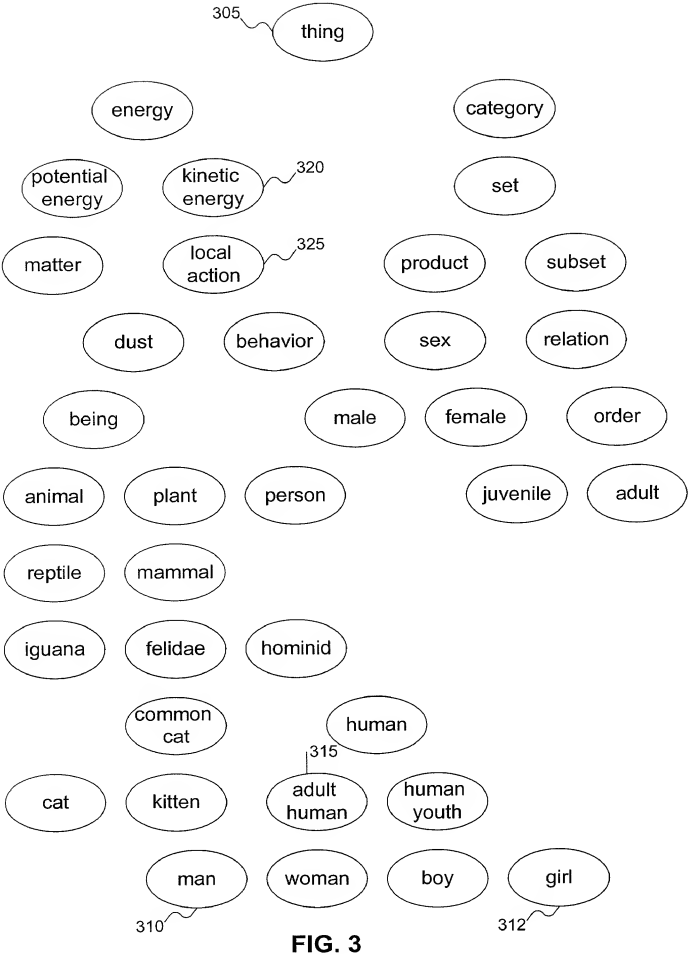

305
thing
energy
category
potential energy
kinetic energy
320
set
matter
local action
325
product
subset
dust
behavior
sex
relation
being
male
female
order
animal
plant
person
juvenile
adult
reptile
mammal
iguana
felidae
hominid
common cat
human
315
cat
kitten
adult human
human youth
man
woman
boy
girl
310
312

FIG. 3

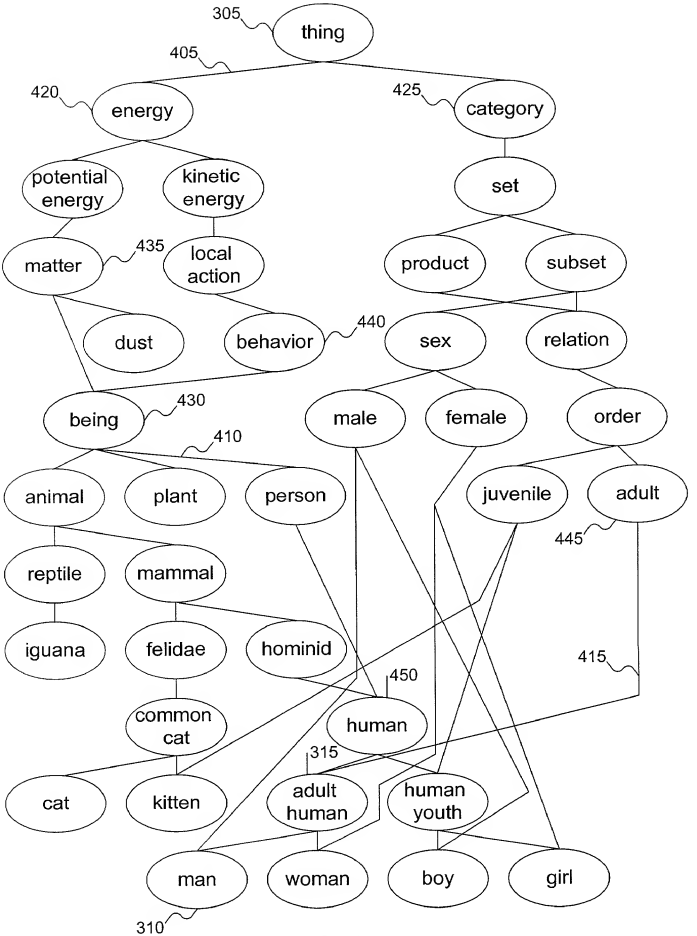

305
thing
405
420
energy
425
category
potential energy
kinetic energy
set
matter
435
local action
product
subset
dust
behavior
440
sex
relation
being
430
male
female
order
410
animal
plant
person
juvenile
adult
445
reptile
mammal
iguana
felidae
hominid
415
450
common cat
human
315
cat
kitten
adult human
human youth
man
woman
boy
girl
310

FIG. 4

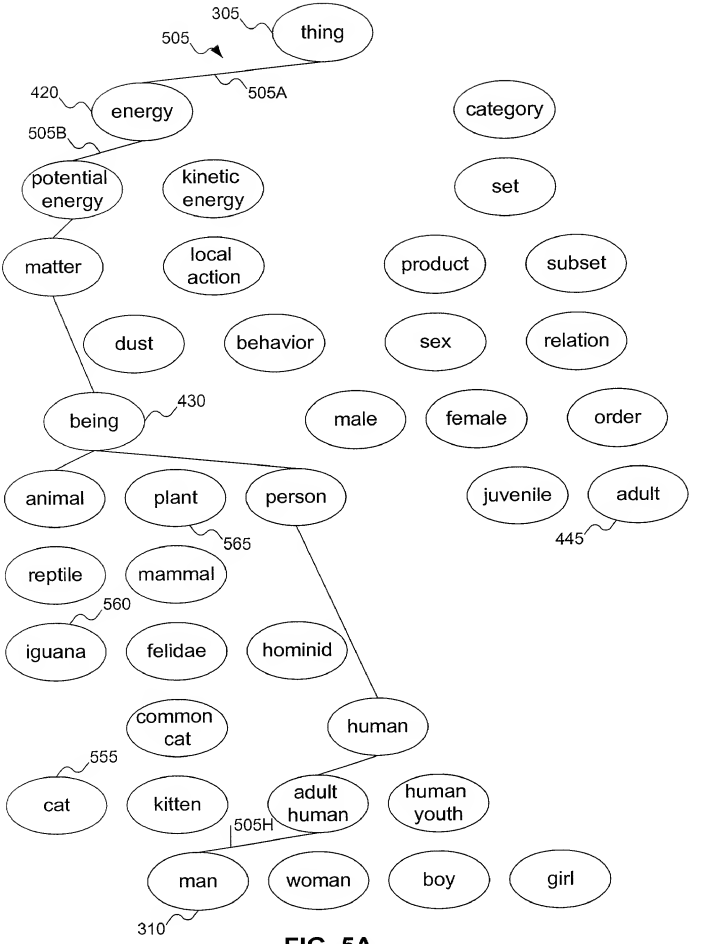

305
thing
505
505A
420
energy
category
505B
potential energy
kinetic energy
set
matter
local action
product
subset
dust
behavior
sex
relation
being
430
male
female
order
animal
plant
person
juvenile
adult
565
445
reptile
mammal
560
iguana
felidae
hominid
common cat
human
555
cat
kitten
adult human
human youth
505H
man
woman
boy
girl
310

FIG. 5A

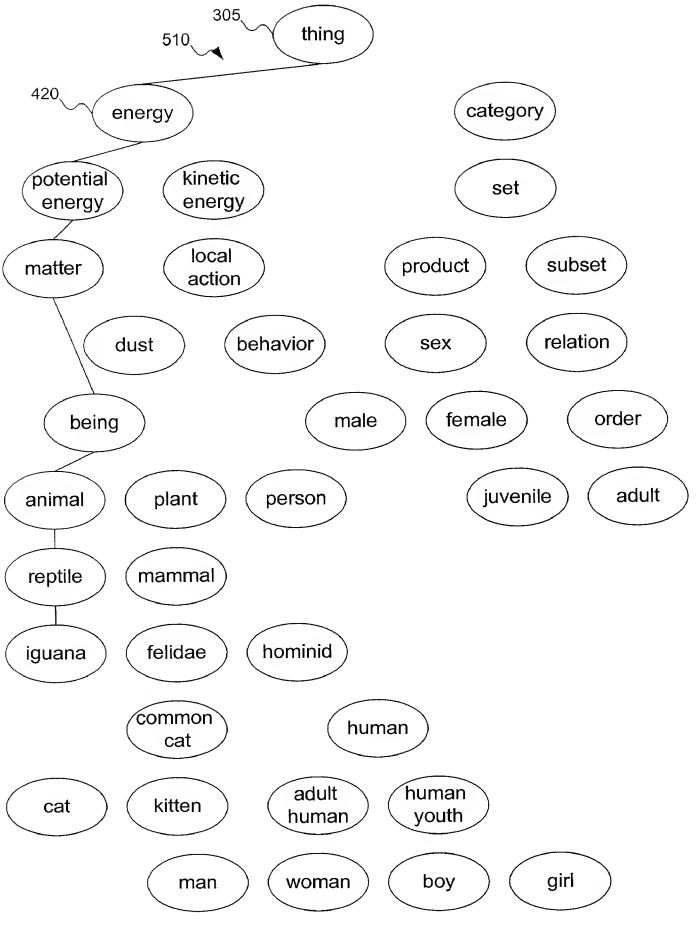

305
thing
510
420
energy
category
potential energy
kinetic energy
set
matter
local action
product
subset
dust
behavior
sex
relation
being
male
female
order
animal
plant
person
juvenile
adult
reptile
mammal
iguana
felidae
hominid
common cat
human
cat
kitten
adult human
human youth
man
woman
boy
girl

FIG. 5B

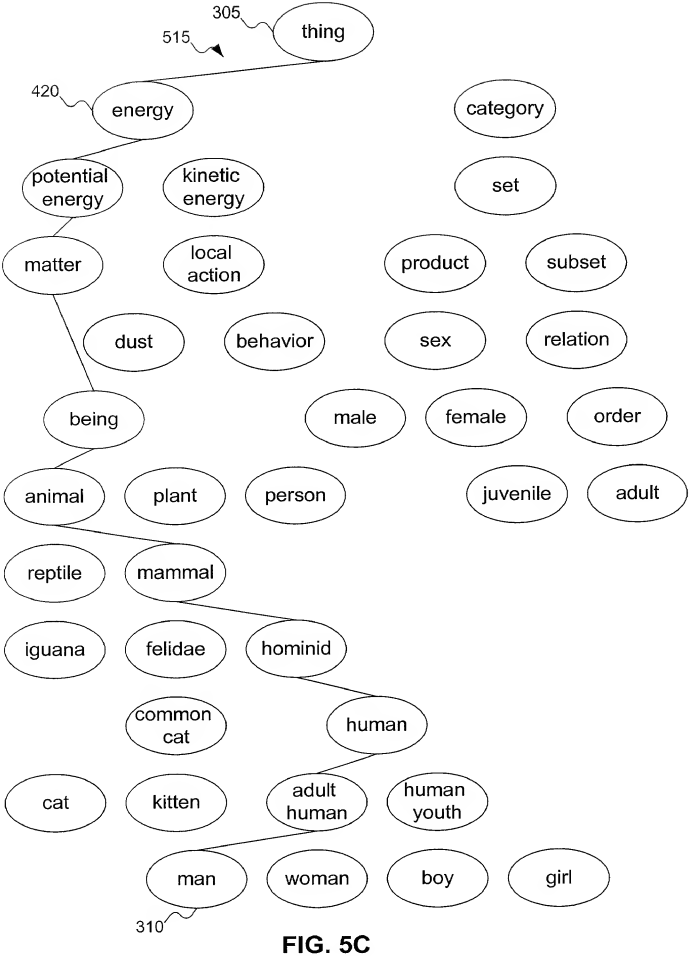

305
thing
515
420
energy
category
potential energy
kinetic energy
set
matter
local action
product
subset
dust
behavior
sex
relation
being
male
female
order
animal
plant
person
juvenile
adult
reptile
mammal
iguana
felidae
hominid
common cat
human
cat
kitten
adult human
human youth
man
woman
boy
girl
310

FIG. 5C

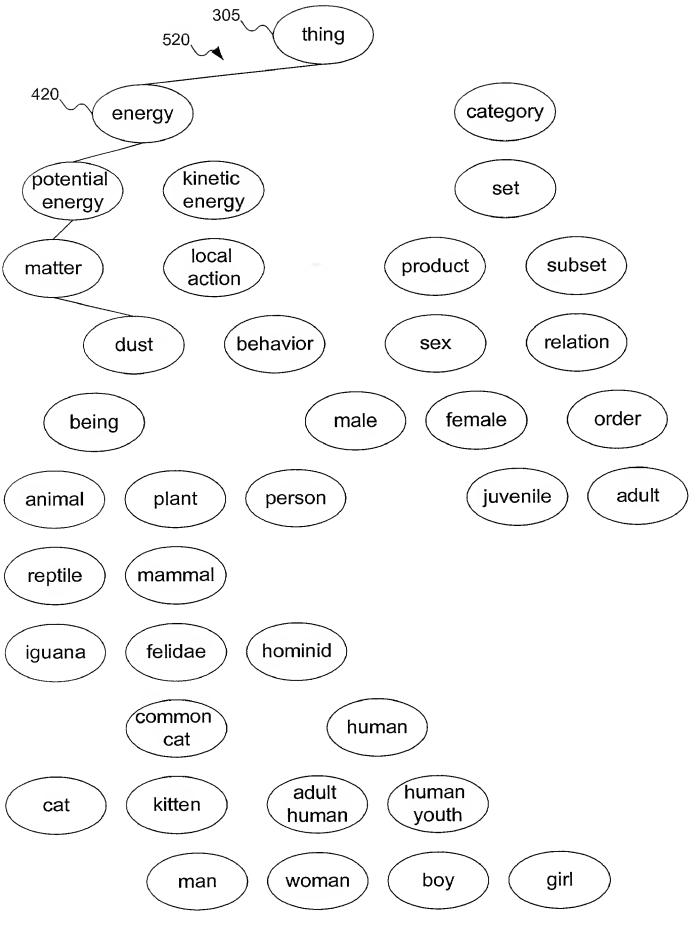

305
thing
520
420
energy
category
potential energy
kinetic energy
set
matter
local action
product
subset
dust
behavior
sex
relation
being
male
female
order
animal
plant
person
juvenile
adult
reptile
mammal
iguana
felidae
hominid
common cat
human
cat
kitten
adult human
human youth
man
woman
boy
girl

FIG. 5D

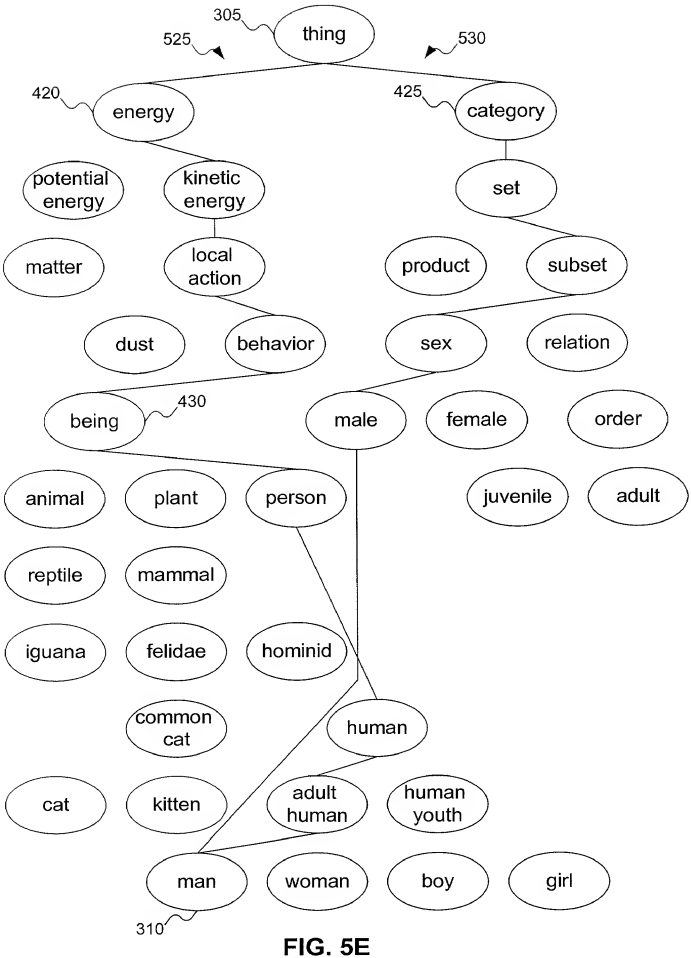

305
thing
525
530
420
energy
425
category
potential energy
kinetic energy
set
matter
local action
product
subset
dust
behavior
sex
relation
430
being
male
female
order
animal
plant
person
juvenile
adult
reptile
mammal
iguana
felidae
hominid
common cat
human
cat
kitten
adult human
human youth
man
woman
boy
girl
310

FIG. 5E

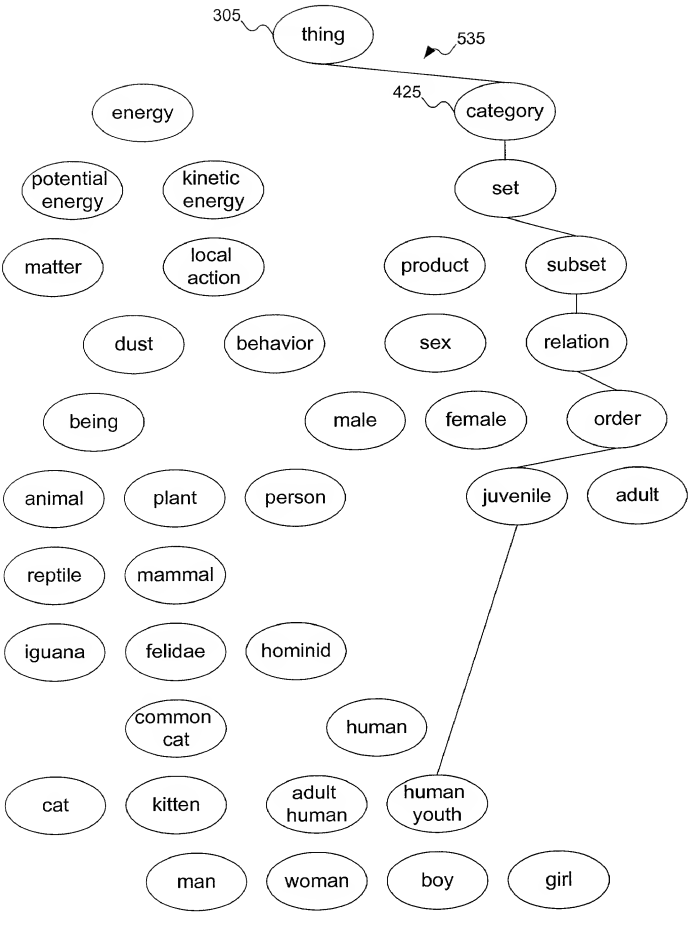

305
thing
535
425
category
energy
potential energy
kinetic energy
set
matter
local action
product
subset
dust
behavior
sex
relation
being
male
female
order
animal
plant
person
juvenile
adult
reptile
mammal
iguana
felidae
hominid
common cat
human
cat
kitten
adult human
human youth
man
woman
boy
girl

FIG. 5F

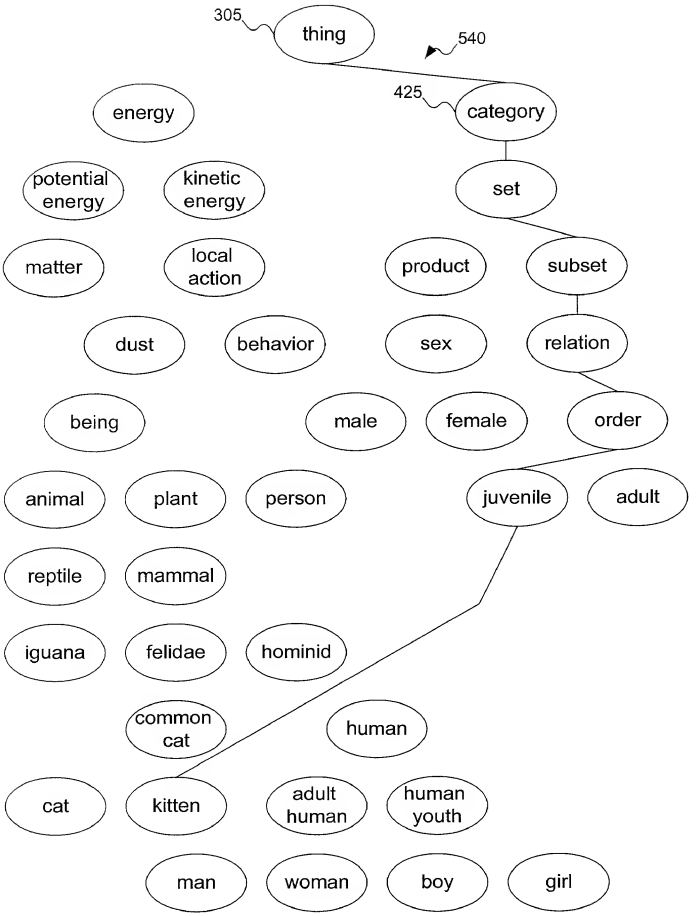

305
thing
540
425
category
energy
potential energy
kinetic energy
set
matter
local action
product
subset
dust
behavior
sex
relation
being
male
female
order
animal
plant
person
juvenile
adult
reptile
mammal
iguana
felidae
hominid
common cat
human
cat
kitten
adult human
human youth
man
woman
boy
girl

FIG. 5G

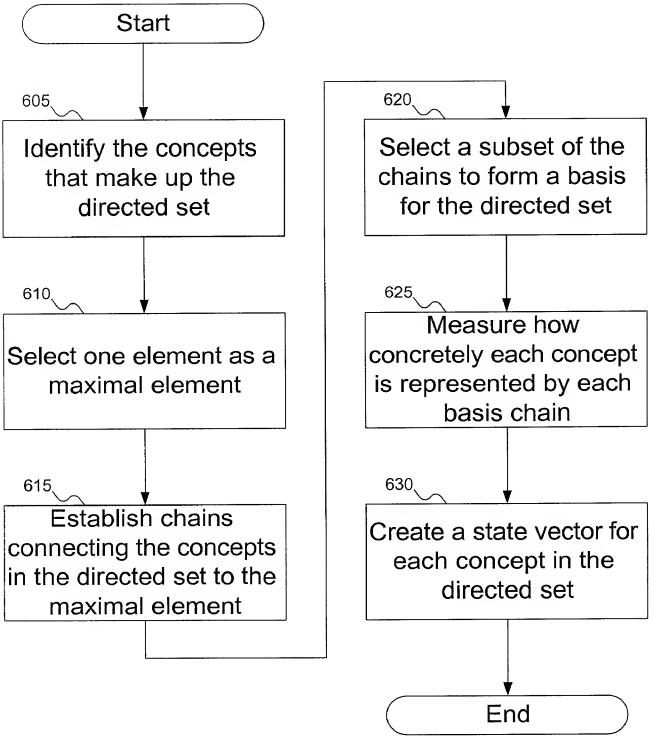

Start
605
Identify the concepts that make up the directed set
610
Select one element as a maximal element
615
Establish chains connecting the concepts in the directed set to the maximal element
620
Select a subset of the chains to form a basis for the directed set
625
Measure how concretely each concept is represented by each basis chain
630
Create a state vector for each concept in the directed set
End

FIG. 6

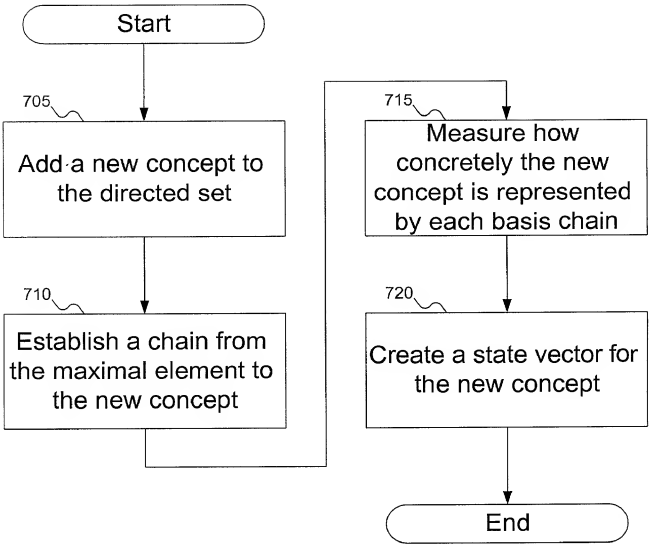

Start
705
Add a new concept to the directed set
710
Establish a chain from the maximal element to the new concept
715
Measure how concretely the new concept is represented by each basis chain
720
Create a state vector for the new concept
End

FIG. 7

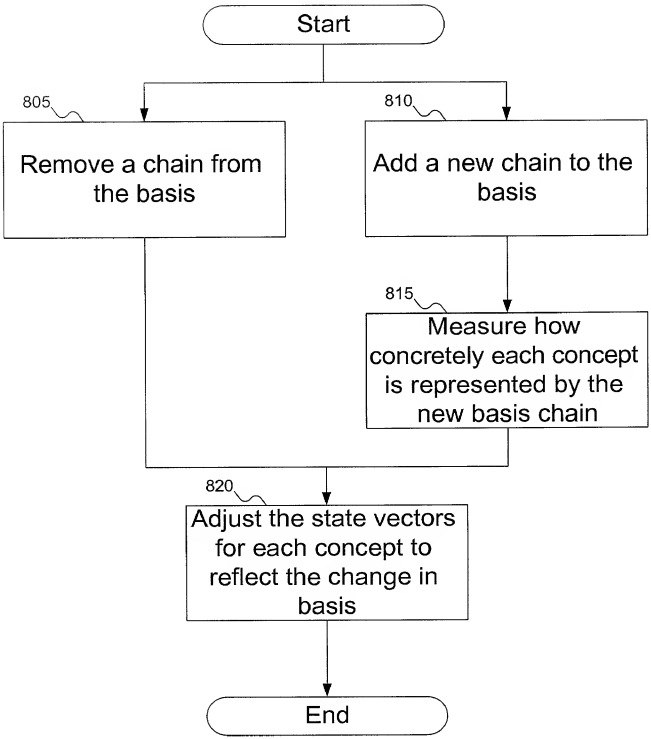

Start
805
Remove a chain from the basis
810
Add a new chain to the basis
815
Measure how concretely each concept is represented by the new basis chain
820
Adjust the state vectors for each concept to reflect the change in basis
End

FIG. 8

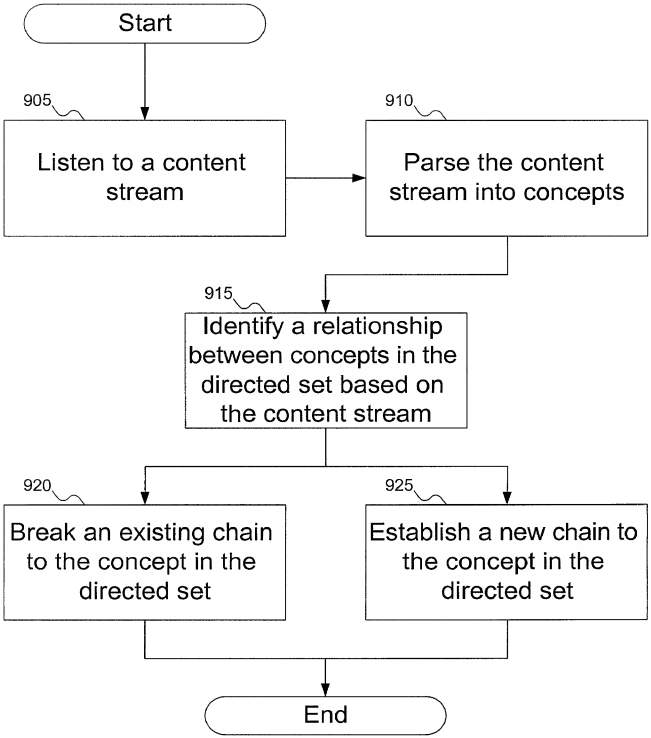
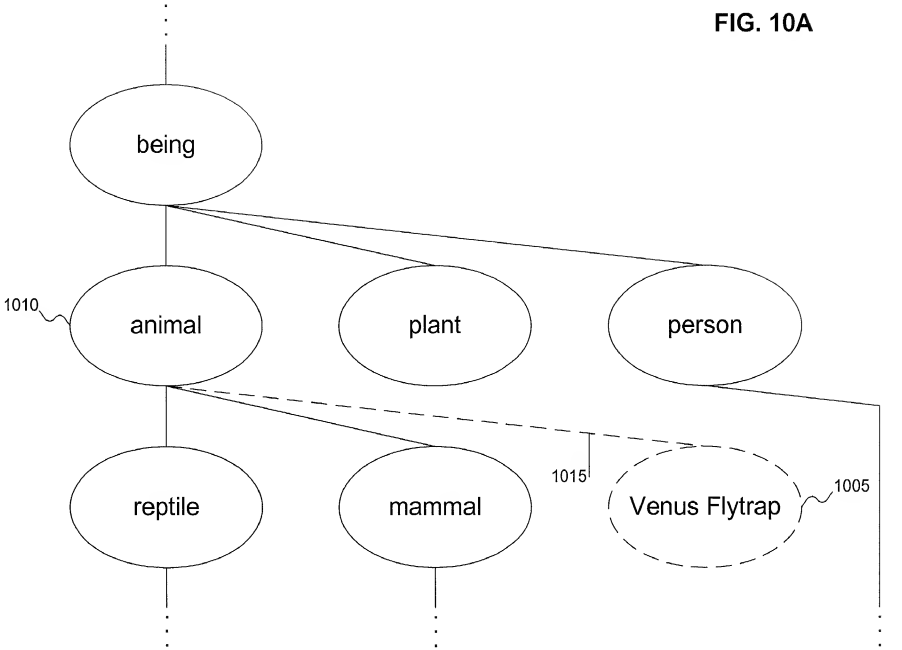

Start
905
Listen to a content stream
910
Parse the content stream into concepts
915
Identify a relationship between concepts in the directed set based on the content stream
920
Break an existing chain to the concept in the directed set
925
Establish a new chain to the concept in the directed set
End

FIG. 9

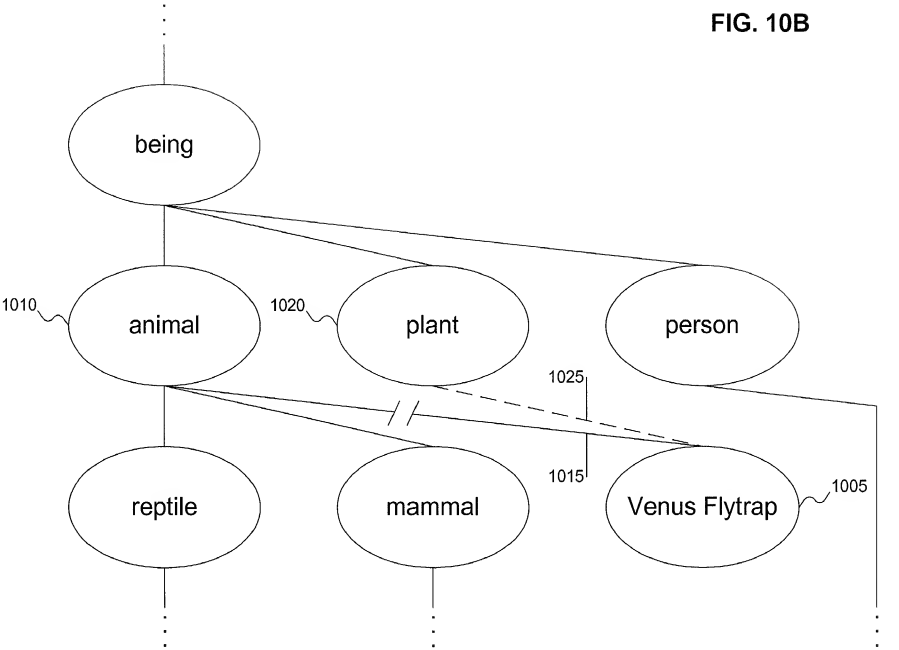
08320-E0E750

FIG. 10A



US 2020/0250509 A1

FIG. 10B



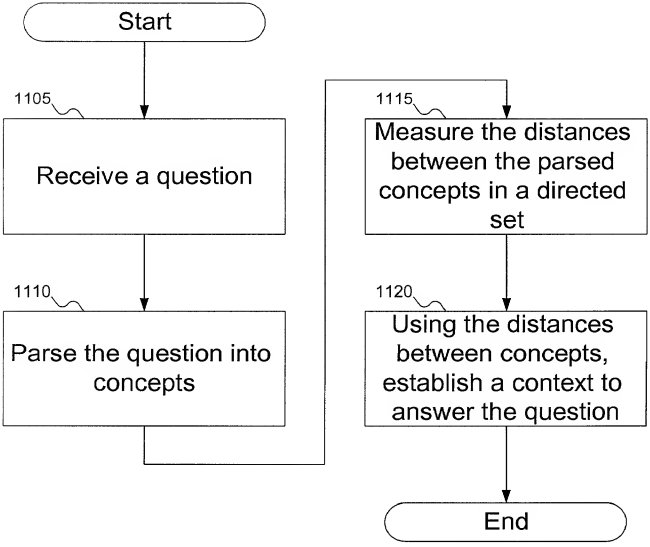

Start
1105
Receive a question
1110
Parse the question into concepts
1115
Measure the distances between the parsed concepts in a directed set
1120
Using the distances between concepts, establish a context to answer the question
End

FIG. 11

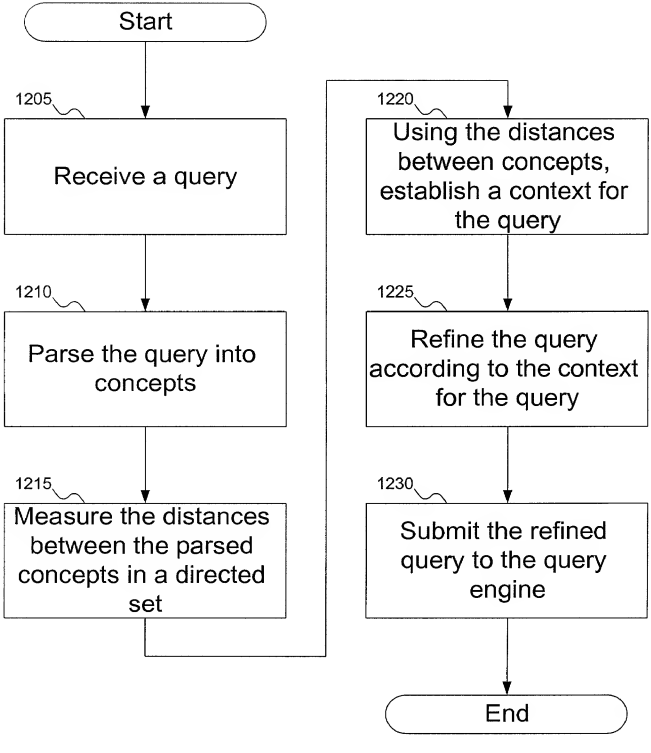

Start
1205
Receive a query
1210
Parse the query into concepts
1215
Measure the distances between the parsed concepts in a directed set
1220
Using the distances between concepts, establish a context for the query
1225
Refine the query according to the context for the query
1230
Submit the refined query to the query engine
End

FIG. 12

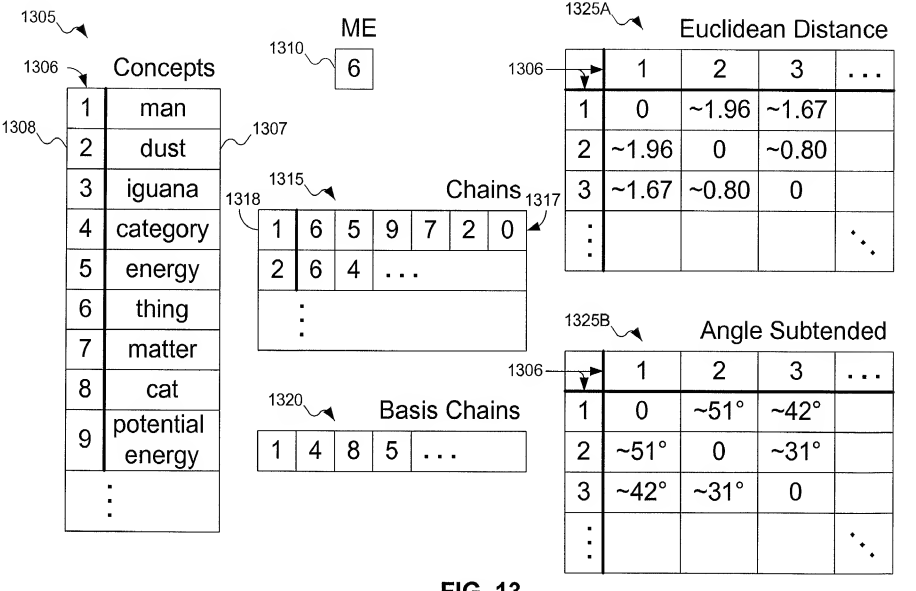

1305
1306
Concepts
1308
1307
1 man
2 dust
3 iguana
4 category
5 energy
6 thing
7 matter
8 cat
9 potential energy
1310
ME
6
1315
Chains
1318
1317
1 6 5 9 7 2 0
2 6 4 . . .
1320
Basis Chains
1 4 8 5 . . .
1325A
Euclidean Distance
1306
1 2 3 . . .
1 0 ~1.96 ~1.67
2 ~1.96 0 ~0.80
3 ~1.67 ~0.80 0
1325B
Angle Subtended
1306
1 2 3 . . .
1 0 ~51° ~42°
2 ~51° 0 ~31°
3 ~42° ~31° 0

FIG. 13

Attorney's Do. No. 6647-3

DECLARATION FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled CONSTRUCTION, MANIPULATION, AND COMPARISON OF A MULTI-DIMENSIONAL SEMANTIC SPACE, the specification of which:

[X] is attached hereto.

[ ] was filed on _____ as Application Serial No. _____

[ ] and was amended on _____ (if applicable)

[ ] with amendments through _____ (if applicable)

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to patentability as defined in Title 37, Code of Federal Regulations, Sec. 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, Sec. 119(a)-(d) of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed: NONE

Prior Foreign Application(s)			Priority Claimed	
(Number)	(Country)	(Day/Month/Year Filed)	[ ] Yes	[ ] No

I hereby claim the benefit under Title 35, United States Code, Sec. 119(e) of any United States provisional application listed below: NONE

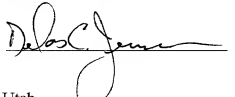
Provisional Application No. Filing Date

I hereby claim the benefit under Title 35, United States Code, Sec. 120 of any United States application(s), or Sec. 365(c) of any PCT International application designating the United States, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT International application in the manner provided by the first paragraph of Title 35, United States Code, Sec. 112. I acknowledge the duty to disclose information which is material to patentability as defined in Title 37, Code of Federal Regulations, Sec. 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application: NONE

09/109,804	July 2, 1998	Pending
(App. Serial No.)	(Filing Date)	(Status -patented, pending, etc.)

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full name of sole or first inventor: Delos C. Jensen

Inventor's signature:  2-16-2000 (Date)

Residence: Orem, Utah

Citizenship: U.S.A.

Post Office address: 635 N. 1250 E.
Orem, Utah 84097

Full name of second co-inventor: Stephen R. Carter

Inventor's signature: _____ 2/10/00
(Date)

Residence: Provo, Utah

Citizenship: U.S.A.

Post Office address: 428 South Nebo Drive
Spanish Fork, Utah 84660

PATENT APPLICATION
Attorney's Do. No. 6647-3

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of: Delos C. Jensen and Stephen R. Carter

Serial No.

Filing Date:

For: CONSTRUCTION, MANIPULATION, AND COMPARISON OF A MULTI-DIMENSIONAL SEMANTIC SPACE

Box Patent Applications
Assistant Commissioner of Patents
Washington, D.C. 20231

ASSOCIATE POWER OF ATTORNEY

Please recognize the following attorneys as my associate attorneys in respect of this application:

Attorney Name	Registration No.
Jerome S. Marger	26,480
Alexander C. Johnson, Jr.	29,396
Alan T. McCollom	28,881
James G. Stewart	32,496
Glenn C. Brown	34,555
Stephen S. Ford	35,139
Gregory T. Kavounas	37,862
Scott A. Schaffer	38,610
Joseph S. Makuch	39,286
James E. Harris	40,013
Graciela G. Cowger	42,444
Ariel Rogson	43,054
Craig R. Rogers	43,888

Please address all telephone calls to Alexander C. Johnson, Jr. at (503) 222-3613.

Please address all correspondence to:

MARGER JOHNSON & McCOLLOM, P.C.
1030 S.W. Morrison Street
Portland, Oregon 97205

NOVELL, INC.

David Bradford
Senior Vice President and General Counsel

NOVELL, INC.
122 East 1700 South
Provo, Utah 84606-6149
(801) 861-4076